





ARTICLE



<https://doi.org/10.1038/s43247-021-00094-x>

OPEN

# Future changes in the frequency of temperature extremes may be underestimated in tropical and subtropical regions

N. Freychet <sup>1</sup>✉, G. Hegerl <sup>1</sup>, D. Mitchell <sup>2</sup> & M. Collins <sup>3</sup>

In a warming world, temperature extremes are expected to show a distinguishable change over much of the globe even at 1.5 °C warming, and in many regions this change has already been detected in observations. Although many studies predict an increase in heat extreme events, the magnitude of the change varies greatly among different models even for the same mean warming. This uncertainty has been linked to differences in land-atmosphere feedback across models. Here we show that a significant constraint for future projections can be based on the ability of climate models to accurately simulate the present day variability of daily surface maximum temperature. An emergent constraint on Coupled Model Intercomparison Project Phase 5 (CMIP5) and 6 (CMIP6) models, applied to ERA5 reanalysis, indicates that the best estimate in hot extreme changes by the end of the century could be worse than previously estimated, mostly for tropical and subtropical regions as well as South and East Asia.

<sup>1</sup>School of Geosciences, University of Edinburgh, Edinburgh, UK. <sup>2</sup>School of Geographical Sciences, University of Bristol, Bristol, UK. <sup>3</sup>College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK. ✉email: [nicolas.freychet@ed.ac.uk](mailto:nicolas.freychet@ed.ac.uk)

Temperature extremes are expected to show a distinguishable change over much of the globe<sup>1</sup>, even at 1.5 °C warming<sup>2</sup>, and in many regions this change has already been detected in observations<sup>3–6</sup>. Change in extremes events remain inconsistent between models, even at a same level of mean global warming<sup>7</sup> in CMIP5<sup>8</sup> and CMIP6<sup>9</sup> models. These extremes have a strong impact on society and can have negative consequences on health<sup>10</sup>, agriculture<sup>11</sup>, or water resources<sup>12</sup>. Daily maximum temperature (TX) is often used to measure heat wave intensity. It is governed by many processes, including solar radiation, heat transport, and sensible, and latent heat flux exchange with the surface<sup>13</sup>. In particular, energy used to evaporate surface moisture can mitigate atmospheric warming and thus TX<sup>14</sup>. At any given location, TX variability tends to be larger under drier surface conditions than wetter conditions. Another way to formulate this idea is that soil moisture deficit (and deficit in other surface humidity variables) can lead to amplified TX (and with it, potentially amplified heat waves) and can explain part of TX variability. There is evidence that many current climate models are too dry under the present conditions<sup>15</sup> and we hypothesize that this amplifies TX variability and with it, heat wave frequency<sup>16</sup>, whereas more accurate models may see this amplification in upcoming decades. We postulate that this could lead to large differences between models in terms of heat wave changes under climate warming. Previous work has indicated a reduction in temperature change uncertainties when using prescribed land heat fluxes<sup>17</sup> and has found a systematic increase in interannual variability of summer temperatures over Europe with models that realistically represent variability<sup>18</sup>. Although our main focus here is on surface heat fluxes, many other processes can also impact TX variability locally (as, e.g., dynamics or aerosols) and model parametrization. Thus, we do not expect to explain systematically the temperature signal from heat fluxes alone.

## Results

**Metrics and techniques.** Many indices of TX can be used to describe hot events (some as defined by the Expert Team on Climate Change Detection and Indices). We chose a simple derived index that can be applied easily at the global scale, namely the number of days above the 98th percentile (TX98p, see “Methods” for detailed computation). We focus only on the warmest season (June to August for North Hemisphere, December to February for South Hemisphere, and all year for the 15°S–15°N tropical area). TX98p indicates for each location the number of days that are considered as extremely hot (relative to the 1995–2005 daily climatology of TX at this location) and we evaluate its change in climate projections (i.e., the change in the number of days above the climatological threshold, see “Methods” for details). We also define a metric to quantify the historical variability of TX at each location,  $\Delta TX$ . This metric indicates, at each grid point and for each calendar day, the distance between mean TX and the 95th percentile of TX in °C.  $\Delta TX$  gives an indication of the temperature difference between a hot day compared to the climatology. It is used to evaluate models against a global reference dataset, the ERA5 reanalysis<sup>19</sup>. This difference between hot and average days has been found to be too high in some climate models<sup>20</sup>. Computation of  $\Delta TX$  implies that we ignore any bias in the mean TX of a model (compared to ERA5) and focus only on TX variability. It is noteworthy that our results are not sensitive to the exact definition of the heat metric. If using another threshold, e.g., the 95th percentile instead of 98th percentile (see Supplementary Fig. 9), results are very similar.

**The role of surface humidity.** Previous studies have shown that soil moisture deficits enhance surface temperature extremes<sup>21–23</sup> as demonstrated, e.g., for European extremes<sup>24</sup>. Here we focus on

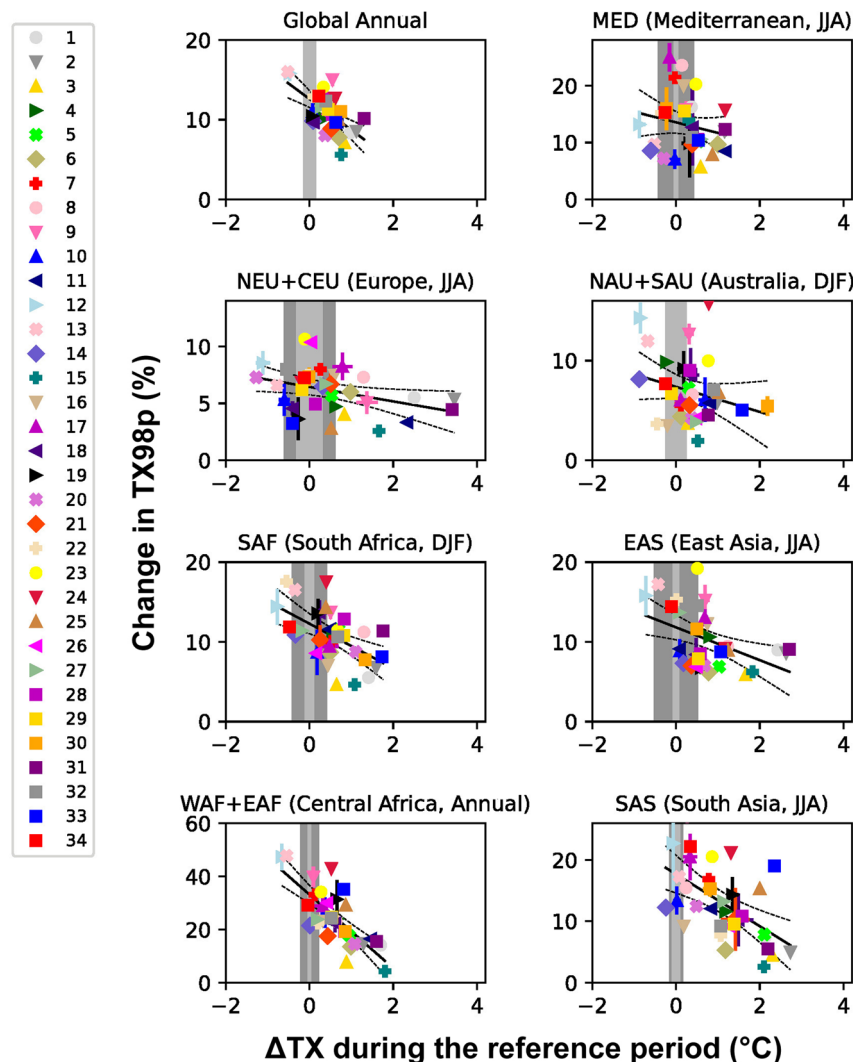
daily temperatures, and due to limited availability of humidity model outputs at daily temporal resolution (especially integrated soil moisture is not available at daily timescales for CMIP5 outputs), we use the surface heat fluxes (latent and sensible) as an indicator of land–atmosphere interaction. Latent heat flux can be a good indicator of evaporation during hot days. We verify that latent heat flux anomalies during days above the 98th percentile exhibit a negative correlation with  $\Delta TX$  over 80% of the land (Supplementary Fig. 1a), i.e., models with high  $\Delta TX$  have less evaporation. Sensible heat flux (Supplementary Fig. 1) has positive correlations almost everywhere (indicating that hot days occur under warmer land surface conditions), except over the driest deserts. In these regions, the air may be so warm that the sensible heat flux is actually reduced. It is also verified that models with the largest change in TX98p have stronger decreases in latent heat flux (Supplementary Fig. 1c) over tropical/subtropical areas and South and East Asia, i.e., they are drying more compared to other models between the present time and the end of the century. This relationship is not observed (or even reverses) for mid-latitudes, corresponding to regions where  $\Delta TX$ –TX98p is not valid as an emergent constraint (EC) and masked in the EC analysis. Other constraints, e.g., related to precipitation change, may be valid there<sup>24</sup>. These results confirm the relationship between surface humidity and TX variability during the baseline period, especially over tropical areas. A noticeable exception are the desert regions of North Africa and Middle East, where we also observe a strong  $\Delta TX$ –TX98p change relationship but it is apparently not related to land-drying processes. In these regions, surface shortwave radiation (related to aerosol, cloud cover, or dynamics) could play a major role to explain  $\Delta TX$  (Supplementary Fig. 2).

## An EC to improve projection in extreme temperature events.

We verify where  $\Delta TX$  is correlated with the TX98p change for different warming targets (end of century or +1.5 and +2 °C warming periods). Over most of land, the relationship between  $\Delta TX$  and TX98p change is negative (Fig. 1) and significant (Supplementary Fig. 3), indicating that in regions with present-day overestimated variance for hot days, the future change in TX98p is smaller on average. Thus, this simple metric is justified, both physically and statistically, to constrain model projections. In the following we apply the EC only where the  $\Delta TX$ –TX98p correlation is significant. This is the case for tropical and subtropical areas and South and East Asia mainly. Mid- and high-latitude show less significant correlations.

The EC methodology requires understanding and accounting for observational and model variability and uncertainties to check consistency<sup>25</sup>. We use the internal variability of a large multi-member historical ensemble (HAPPI<sup>26</sup>), which is forced with observed sea surface temperatures to estimate the uncertainty in the  $\Delta TX$  variability associated with weather variability at each location (model internal variability/error). As an estimate of observational error, we use the difference between ERA5 and the blended satellite surface CHIRTS<sup>27</sup> dataset, both of which provide high-resolution daily output for TX. We then consider these two pieces of information as an uncertainty range for  $\Delta TX$  based on ERA5 and to evaluate when models fit within this range (with multi-member models having narrower uncertainty, see “Methods”) and select only these models to simulate future change in heat wave metrics.

**Regional and global constraint.** We find that changes in TX98p are larger than estimated by an unconstrained ensemble over a large part of tropical and subtropical regions when using  $\Delta TX$  to constrain climate projections by selecting at the regional scale (see

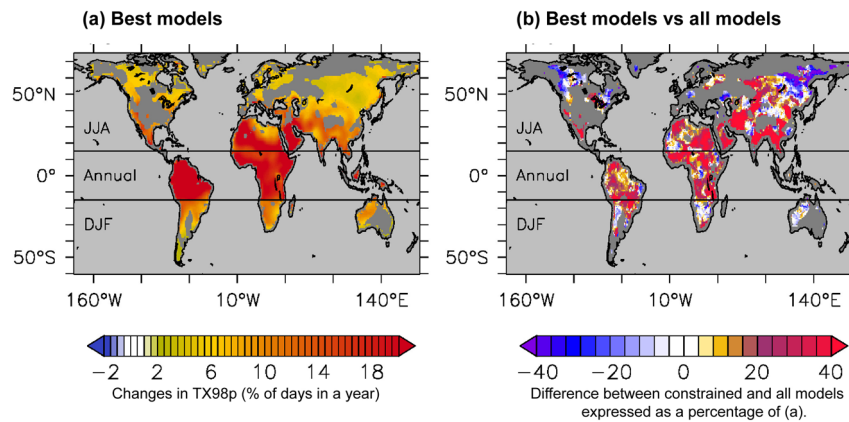


**Fig. 1 Relationship between  $\Delta$ TX and projected change in TX98p in selected regions.** The figure shows for each CMIP5 and CMIP6 (square symbols) model the change in the ensemble average frequency of hot days (TX98p, y axis, in % of days) in the future (last decade of RCP45 and SSP245) compared to the present period (1995–2005) plotted against a variability metric for daily maximum temperature ( $\Delta$ TX) during the historical period (x axis, in  $^{\circ}$ C).  $\Delta$ TX measures the difference between daily TX95p and mean TX in a model compared to that observed. The solid black line is the linear regression between  $\Delta$ TX and TX98p, and dashed black lines show the 95% confidence interval for the slope line. Light gray shading represents the  $\Delta$ TX difference between ERA5 and CHIRTS (observation uncertainty), and dark gray shading shows uncertainties estimated from the HAPPI ensemble (model internal atmospheric variability) combined. Acronyms refer to AR5 region definitions and numbers refer to models in Supplementary Table 1.

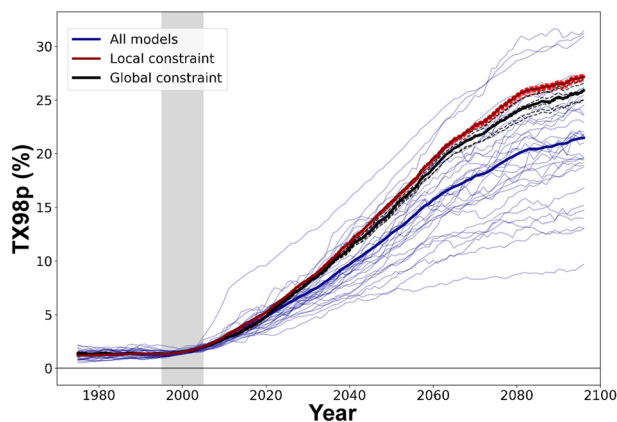
“Methods”) the models within the observed constraint (Fig. 2). Africa, South and Central Asia, and South America have a particularly strong signal, locally above 50% increase in the number of exceedances of the 98th percentile, i.e., twice as many hot days as in unconstrained predictions. This means models that represent  $\Delta$ TX more accurately during the baseline period (and thus, hypothetically, humidity feedbacks) tend to warm hot extremes faster compared to the other models. Similar relationships are found for all climate warming targets (Supplementary Figs. 10 and 11), although the area with significant correlation is reduced for a 1.5  $^{\circ}$ C target. Thus, the influence of our EC persists through different warming targets and is confirmed robustly by several sensitivity tests. It is also able to improve the prediction of any climate model’s future change in the heat wave metric based on the ensemble, passing a key test for ECs (perfect model test; see ref. <sup>25</sup> and “Methods”). As an uncertainty, we note that EC results are the strongest where only a few models are selected as realistic (Supplementary Fig. 4a). Thus, the magnitude of the

amplification of heat wave projections may be less robust over these areas, as it relies on a more limited number of models. Applying an EC based on global mean  $\Delta$ TX (i.e., selecting single models based on a global mean relationship only) leads to slightly weaker, but still valid, amplification with seven models selected as more realistic (Fig. 3). Using a regional constraint to select the best models at regional scale seems more appropriate, as no model is considered realistic in hot day variability everywhere (Supplementary Fig. 4).

The constrained TX98p signal (either by the local or global method) suggests that the level of increase in frequency of hot days that was previously estimated to occur by the end of the century could be reached by 2060 instead, i.e., 40 years earlier. All these results are verified to be independent of model selection by performing sensitivity tests where one model is removed randomly from the ensemble (Fig. 3). The regional constraint remains highly skillful in this sensitivity test. The global constraint is still consistent but slightly more sensitive to model



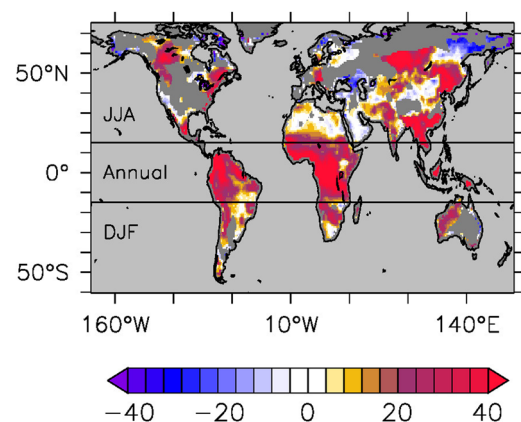
**Fig. 2 Implication of regional emergent constraint for future change in extremes.** **a** Ensemble mean (all CMIP models) change in the frequency (by the end of the century) of hot extremes TX98p per degree warming compared to the baseline 1995–2005 period, expressed as a percentage of days (+X% means an extra X% of days each year will be above the 98th percentile, see “Methods”). **b** Difference in TX98p projections between models that reproduce the observed constraint locally and all models, expressed as a percentage of the change in **a**. **a**, **b** Results are only shown where the correlation between TX98p and  $\Delta$ TX is significant (see Supplementary Fig. 3).



**Fig. 3 Global evolution of hot extremes in unconstrained and constrained ensembles.** Time series of global mean TX98p (%) for the mean (thick solid blue line) and individual models (thin blue lines) of all CMIP5 (using RCP4.5 from 2006) and CMIP6 (using SSP245 from 2015) ensembles. Regionally constrained model results are shown in red. A 9-year running mean of TX98p is first applied at each grid point then globally averaged to obtain yearly global mean values. The solid black line shows the mean of a sub-ensemble (seven models) where EC is based on globally averaged  $\Delta$ TX (instead of applying EC regionally). Gray shading highlights the baseline period used to compute the TX98p threshold and the constraint. Red (and black) dashed lines show sensitivity tests where one model is removed from the ensemble before computation of regional (and global) emergent constraints (test repeated for each model of each ensemble). For each model and each year, TX98p is first normalized by the change in mean  $T_{as}$  then linearly scaled to the multi-model ensemble mean increase in  $T_{as}$  (see “Methods”; note EC results are not sensitive to this, Supplementary Fig. 6).

selection (due to the small size of  $n$  models that fall near the global variability uncertainty range). Similar spatial patterns of increased frequency in extremes to those shown for the regional constraint are also found if applying the global constraint (Fig. 4), except over desert areas where the signal is not present. This again suggests distinct processes between desert regions and other parts of the globe.

**Physical constraints.** To further evaluate the EC, we evaluate the physical mechanism linking change in TX98p and land drying. The relationship between the future change in TX98p and latent



**Fig. 4 Implication of global emergent constraint for future change in extremes.** Similar to Fig. 2b but for emergent constraint applied globally for the end-of-century warming (corresponding to solid black line in Fig. 3; with seven models selected by the EC).

heat flux is overall negative where the EC is significant and has the strongest influence on results (Supplementary Fig. 1), indicating larger temperature variability for weaker fluxes (drier soils), supporting our hypothesis. Larger sensible heat flux (heat transfer from land to atmosphere) during hot extremes (Supplementary Fig. 1c) also highlights the importance of land–atmosphere interaction during hot days for our EC. We find a decrease in latent and an increase in sensible heating that correlates with an increase in the frequency of hot extremes. We note that over some regions constrained models do not indicate an increase in TX98p, especially over northern parts of America and generally in mid-latitudes. These correspond to areas with weak correlation between  $\Delta$ TX and TX98p at present. Other processes may be more dominant in these regions (including change in rainfall and atmospheric dynamics) and drying of soil may be not a factor in high latitudes. In addition, permafrost land–atmosphere exchanges and humidity processes are different there.

## Discussion

Our results indicate that climatological bias in the difference between hot and average days in climate models leads to an



underestimate of the frequency of unusually hot days in the future over many low latitude regions. If the EC is based on the latent heat flux (i.e., the ability of models to reproduce correctly the land–atmosphere humidity exchanges) rather than temperature variability, results support findings over regions where the EC is statistically strongest (tropical regions and South-East Asia). This suggests that changes in hot extremes are related to land surface humidity processes. Results over other regions are less clear, suggesting that other physical processes may dominate the changes in temperatures variability there. When using the regional constraint, the limited number of models selected over some areas may question the robustness of the actual EC signal. However, results using a global constraint are consistent with regional constraints and support the key findings of the latter methodology.

We further note that we focus on daily dry-bulb temperature as metric for hot extremes. In the tropical and subtropical regions, where our constraint is most valid, dangerous heat stress for humans specifically will also be influenced by humidity, which is not considered here. High humidity can lead to deadly conditions when combined with high temperatures and could be a major threat in the future<sup>28,29</sup>. As our EC is physically linked to heat and humidity exchange between land and atmosphere, the wet-bulb temperature signal could be investigated with a similar methodology.

## Methods

**Definition and computation of indices.** Our analysis focuses on TX extremes (TX98p). We define TX98p as the number of days above the daily climatological 98th percentile during the warm season (see below). The latter is computed for each location and each calendar day by pooling together all days within a  $\pm 15$  days window of this calendar day during the 1995–2005 period and selecting the 2% highest values.

We also define a metric,  $\Delta TX$ , to evaluate the variability of TX during the baseline period over the warm season. It is defined by first calculating the mean and 95th percentile of the temperature distribution for each calendar day at each location (by pooling 15 days around each calendar day as for 98th percentile described above). The distance between the 95th percentile and the mean gives an indication of TX variability for each day and each location. It is computed for each model ( $\Delta TX_{\text{model}}$ ) and for the reference dataset ( $\Delta TX_{\text{ref}}$ ; the ERA5 reanalysis) and the difference between the two defines our metric:  $\Delta TX = \Delta TX_{\text{model}} - \Delta TX_{\text{ref}}$ . Figure 5 illustrates the computation of these indices.

We only focus on the warm season, when hot extremes are likely to happen (June–August for the Northern Hemisphere, December–February for Southern Hemisphere, and all year for the 15°S–15°N tropical areas). Positive values of  $\Delta TX$  mean a model overestimates the TX variability compared to the reference (i.e., it overestimates high values of TX), negative values indicate an underestimate. For the metric, we choose the 95th percentile to ensure reasonably good sampling of the variability across the base period (as it is used to constrain models), whereas for future changes we focus on the 98th percentile, which correspond to more extreme values. We verified that EC results are not very sensitive to the choice of threshold by doing a sensitivity test using the 95th percentile as threshold instead 98th (Supplementary Fig. 9).

Each index is computed individually for each model and each ensemble member on their native grid. Results are then interpolated on a common 1° grid before being averaged across all models. As temperature extremes are relatively large scale and grids vary only between 1° and 2.5° latitude/longitude across models, results should not be sensitive to the order of operation.

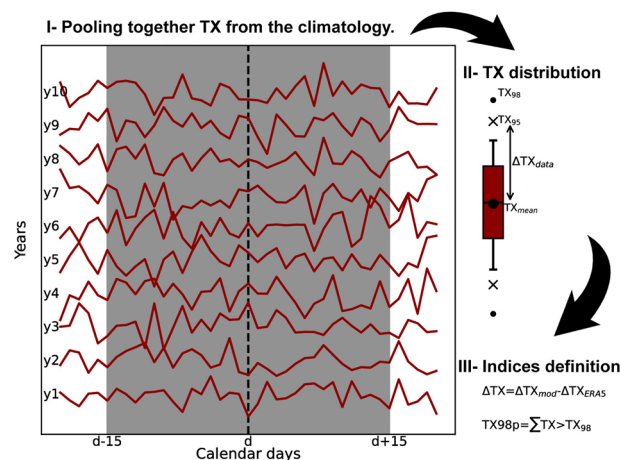
## Datasets

**CMIP models.** An ensemble of 27 individual models from CMIP5<sup>8</sup> and 7 from CMIP6<sup>9</sup> is used. Some only have a single member available while some provide a multi-member ensemble. In the latter case, multi-member results are always computed individually and then averaged to provide one mean result for a single model. We consider a reference period as the historical 1995–2005 decade (being the last decade of CMIP5 historical forcing). Climate projections are investigated using the RCP4.5<sup>30</sup> and SSP245<sup>31</sup> pathways for CMIP5 and CMIP6 models, respectively. Both scenarios are expected to be similar, although each model leads to different mean temperature increases (Supplementary Fig. 5).

Three climate projection targets are considered as follows:

End-of-century, by selecting the 2091–2100 decade for each model.

A 1.5°C and +2°C warming above the pre-industrial mean. For these two, we follow a similar approach as in previous study<sup>32</sup> and select for each member of each model the first decade when the average atmospheric surface temperature (Tas) of



**Fig. 5 Schematic figure to illustrate the computation of indices.** **I** For a dataset, data (model or ERA5), daily maximum temperatures (TX) around a calendar day  $d$ , and during the 10 climatological years are pooled together. **II** The distribution of these temperatures is used to compute the mean ( $TX_{\text{mean}}$ ) and the 95th ( $TX_{95}$ ) and 98th ( $TX_{98}$ ) percentiles. The difference between  $TX_{95}$  and  $TX_{\text{mean}}$  is computed ( $\Delta TX_{\text{data}}$ ). **III**  $\Delta TX$  and TX98p are then computed from the previous results.

each year of the decade is above the corresponding threshold (Supplementary Fig. 5). As we use 1995–2005 as a baseline, the actual threshold (relative to the baseline) is chosen as +0.7°C and +1.2°C for targets +1.5°C and +2°C above pre-industrial, respectively, as in the HAPPI experiment design<sup>26</sup>. Although the exact definition of these levels can be sensitive to how the baseline period is defined<sup>33</sup>, for this work the main point is that each model or member should reach a similar magnitude of warming. A few members and models do not meet the condition for reaching the +2°C target before the end of the century. For these cases, we select instead the last projection decade 2091–2100. If the mean increase in Tas over this decade is above the threshold (+1.2°C), then we keep the model or member. Otherwise we do not include it in the analysis for this projection target. This leads us to discard four members.

For each climate projection target, results of each member or model are normalized by their respective mean change over the decade (relative to our baseline) in Tas and then averaged to provide ensemble mean results. Thus, for the three specific target projections, all results are shown normalized to an overall warming of +1°C above the baseline.

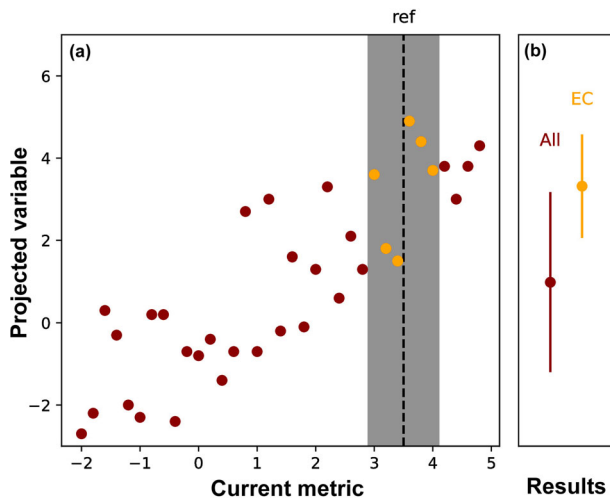
When showing time series of TX98p (Fig. 3), normalized results for individual models are re-scaled to the multi-model ensemble mean increase in temperature for each year (to include more explicitly the global warming trend). We tested the sensibility of the results by using raw results (without normalization) for each model but both methods lead to very close results in terms of EC amplification (Supplementary Fig. 6), although raw results have larger uncertainties. Thus, we largely focus on normalized results in the body of the paper.

For most of the models we could get daily TX data for both historical and projection periods while heat flux daily data are more limited. Supplementary Table 1 provides details about outputs used for each variable.

**HAPPI ensemble and  $\Delta TX$  uncertainty.** To evaluate the model uncertainties on  $\Delta TX$  during the baseline period, we use several atmosphere only model simulations driven with observed SST patterns from the historical HAPPI ensemble<sup>26</sup>. Each model provides daily output for the 1995–2005 decade. We select 5 models with a 100 or more members and compute  $\Delta TX$  for each member (same method as for CMIP models). Then, using the internal variability of each model (multi-member ensemble SD,  $\sigma$ ), we estimate  $\Delta TX$  uncertainties for each location and calendar day (Supplementary Fig. 7). One model has a mean bias that is much larger than other models (CanAM4) and we thus exclude it (although, as its internal variability is close to other models, we found similar results when including it too). For other models, the  $\Delta TX$  internal variability is consistent, so we use the mean of four remaining model variabilities (i.e., averaging the four internal standard deviations, STDs) as a measure of  $\Delta TX$  uncertainties ( $\sigma_{\text{HAPPI}}$ ).

The sensitivity of this choice is also tested by using individual model STD instead of ensemble mean (Supplementary Fig. 8). It shows that results stay consistent for each case. We note that the uncertainty so described is that of atmospheric variability only. However, both the HAPPI ensemble and the ERA5 reanalysis are driven by the same SSTs hence this choice is conservative to characterize observational uncertainty.

Internal variability in the climate models used is reduced by ensemble averaging. To take into account the specific number of members for each individual model, the uncertainty between observation, OBS, ( $\sigma_{\text{OBS}}$ , see “ERA5” section



**Fig. 6 Schematic illustration of the emergent constraint (EC)**

**methodology.** The projection of a given variable (TX98p in our analysis) is first estimated by an ensemble of individual models (a, red and yellow circles), providing a mean and uncertainty in the projection (right panel, red symbol). A metric ( $\Delta TX$  in our analysis) is then computed in both models and a reference (ref) during the historical period. It must be verified that this metric is reasonably correlated to the projected variable and that a physical link can be explained. If so, models within the error range (gray shading, corresponding to the observational error and model internal variability) are selected as good models (yellow circles). The projection using only these selected models (b, yellow symbol) is computed and can be compared to the first estimated projection. An efficient EC will usually provide a more accurate projection (reduce uncertainties) and eventually a different estimate of the mean forecast.

below) and models is expressed as the quadrature sum:  $(\sigma_{OBS}^2 + (\sigma_{HAPPI}^2/N))^{1/2}$  with  $N$  the number of members of a model. When the absolute value of  $\Delta TX$  fits within that range then a model (eventually the multi-members ensemble mean) is considered as consistent with OBS.

**ERA5.** The ERA5 reanalysis<sup>19</sup> is available for the full satellite observation period (1979–present). It provides hourly timescales data at  $0.25^\circ$  resolution on a reduced Gaussian grid, from which we computed daily TX for the 1995–2005 period.

We evaluated the variability of TX in ERA5 against two dense regional observational datasets (Supplementary Fig. 13): a network of 756 homogenized station measurements for China, provided by the Chinese Meteorological Administration<sup>34</sup>, and gridded  $0.25^\circ$  E-OBS v19.0 dataset for Europe<sup>35</sup>. Chinese observations are first gridded on the same regular grid as ERA5 by linear interpolation. Although the TX variability tends to be weaker in ERA5 than in observations, differences are within the range of uncertainties estimated from the HAPPI ensemble variability (Supplementary Fig. 7) for both regions; hence, we consider ERA5 sufficient.

Another observational dataset is used to evaluate ERA5 globally: The Climate Hazards Center Infrared Temperature with Stations CHIRTS<sup>27</sup> (Supplementary Fig. 14). Some differences are visible locally but are within the range of HAPPI variability. The difference between ERA5 and CHIRTS is used as an observation error ( $\sigma_{OBS}^2$ ) and added (in quadrature) to HAPPI variability when applying the EC (see above).

**EC method.** To decrease model projection uncertainties, model weighting based on individual model performances can be applied, providing an accurate knowledge of the single model skill<sup>36</sup>. This remains a challenge in the presence of regional dynamics<sup>37</sup>. Here we use an EC method based on model selection. This method is illustrated by Fig. 6.  $\Delta TX$  is our target metric, i.e., we select models that are able to reproduce the width of the TX distribution (the distance between the 95th percentile and the median) within observational and natural variability uncertainty, and select those models for prediction. It has already been demonstrated that TX variability is a useful indicator to weight models and reduce projection uncertainties<sup>38</sup>. To do this here, CMIP models are evaluated against ERA5 during the 1995–2005 period and selected to agree with it within atmospheric internal variability. We use variability from the HAPPI ensemble to characterize variability uncertainty for better sampling, combined with an estimate of observational uncertainty (see above). Models (ensemble mean in case of multi-members model) within the range of error (as described above) are considered reasonably realistic

and selected for use in the constrained climate projections. Comparing constrained against unconstrained ensemble projections provides an estimate of the potential current bias in climate forecasts.

Constraints may be applied using global or regional processes<sup>25</sup>. Here we use a regional constraint to take advantage of model spatial information. We first apply a spatial smoothing of  $5^\circ$  on  $\Delta TX$  over land (to improve sampling and avoid spatial discontinuity; although results are not very sensitive to the scale of the spatial smoothing, see sensitivity test discussed below) then select the models that comply with the constraint within uncertainty at each grid point. Over most of the regions, the number of selected models is between 5 and 10, except in central Africa where it is below 5. This is mainly due to very narrow observational variability over this region (Supplementary Figs. 4 and 7). Most of the models contribute to the projection over some part of the land. Applying an EC at a global scale instead (Figs. 3 and 4) leads to similar patterns with slightly weaker amplification of future projections.

We also tested the sensitivity of EC results to different choices of uncertainty around the observational distribution and different spatial smoothing (Supplementary Fig. 8). Using narrower (wider) range of variability leads to slightly different results with less (more) models selected. This corresponds to a noisier but more intense signal for narrow smoothing (and opposite for larger smoothing). However, global patterns are still consistent with main results. Weaker spatial smoothing ( $3^\circ$ ) leads to slightly noisier results, while using a larger smoothing area ( $11^\circ$ ) leads to a large masked area (because we use only land grid points or alternatively to large variation in actual applied smoothing). Thus,  $5^\circ$  smoothing is a good compromise.

Following previous recommendations<sup>25</sup>, we first confirm the strong statistical relationship between  $\Delta TX$  and TX98p (Fig. 1 and Supplementary Fig. 3). We then use a resampling method (by removing randomly a model from the ensemble) to test the robustness of the constraint (Fig. 3). Finally, the physical mechanism hypothesis linking land–atmosphere interactions,  $\Delta TX$  and TX98p is evaluated (Supplementary Fig. 1) and a perfect model test is conducted (see below).

**Validity of the EC method.** First, to avoid selection bias<sup>39</sup> and to verify that results are reflective of a physical constraint and roughly independent from the choice of the metric this constraint draws on to determine our EC, we performed a similar analysis with a set of other indicators, all potentially related to Tmax variability, as follows: the interannual variability of Tmax, the diurnal temperature range (DTR) variability, the surface latent heat flux variability, and the surface sensible heat flux variability. Indeed, previous studies have shown the link between heat fluxes variability and temperature variability<sup>40</sup>. All indicators are computed in similar way to  $\Delta TX$  (except that the fifth percentile is used for latent heat flux as it corresponds to drier conditions). Results of EC on TX98 using these indicators are shown in Supplementary Fig. 12. All indicators lead to very similar results to those using TX variability. It confirms that the EC applies to most of the tropical/subtropical areas and South and East Asia (i.e., humid regions). They also confirm that mid- and high latitudes do not show similar results (as it is already the case when using  $\Delta TX$ ); thus, different processes are involved in these areas. The main difference occurs over North African and Middle Eastern arid regions when using DTR as a constraint, with a decrease in projection compared to the increase using our EC. This may be due to the fact that DTR variability is also related to Tmin which is expected to have a larger variation over dry regions. Thus, large DTR variability may be an indication of models cooling down too quickly during the night or warming up too quickly during the day, which makes results less reliable here and DTR less suitable for an EC. Results may also be influenced by shortwave radiation and with it clouds (Supplementary Fig. 12).

Second, to verify the validity of our regional constraint, we used a perfect model test as used in previous studies<sup>41</sup>. We select a model as a reference instead of observations, and apply the EC using this new reference, and then compare the projection in TX98p (after excluding the reference model from the ensemble projections) between the constrained models versus all models to the target of prediction (the prediction by the selected model). We first chose the model which showed the largest fraction of grid points consistent with ERA5 (model 34 in Supplementary Table 1, IPSL-CM6A-LR). Results of this test are shown in Supplementary Fig. 15. It clearly indicates that the error in the TX98p projection for this model (difference between the reference model and other models) is reduced for constrained models in the tropical and subtropical areas (i.e., where we also see the strongest signal with EC using ERA5). It is worth noting that even over desert regions, the error is reduced despite the less clear mechanism there. We also tested the same method using a model that compares less favorably to the observed constraint as target (model 1 in Supplementary Table 1, BCC-CSM1) and results are shown in Supplementary Fig. 16. Improvements are also very clear, indicating that “bad models” tend to attract each other and are consistent in their hot extreme projections. Finally, we repeated the process for all individual models and confirmed that errors are always reduced for constrained ensembles versus all models mean (Supplementary Fig. 17). This is a powerful confirmation of the validity of our EC<sup>25</sup>.

## Data availability

The authors declare that all data that support the findings in the main article are available. All model data are publicly accessible via the Earth System Grid Federation node (<https://esgf-node.ipsl.upmc.fr/>). ERA5 data can be downloaded from ECMWF website (<https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5>). CHIRTS data can be downloaded from <https://www.chc.ucsb.edu/data/chirtsdaily>.

## Code availability

Scripts used to generate the main results and other data and codes that support the figures in the Supplementary Information are available from the corresponding author on request.

Received: 3 August 2020; Accepted: 9 December 2020;

Published online: 04 February 2021

## References

- Herring, S. C., Hoerling, M. P., Kossin, J. P., Peterson, T. C. & Stott, P. A. Explaining extreme events of 2014 from a climate perspective. *Bull. Am. Meteorol. Soc.* **96**, S1–S172 (2015).
- Baker, H. S. et al. Higher CO<sub>2</sub> concentrations increase extreme event risk in a 1.5 °C world. *Nat. Clim. Change* **8**, 604–608 (2018).
- Seneviratne, S. I. et al. *Changes in climate extremes and their impacts on the natural physical environment*. Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the Intergovernmental Panel on Climate Change. 109–230 (2012).
- Bindoff, N. L. et al. in *Climate Change 2013: The Physical Science Basis. IPCC Working Group I Contribution to AR5* (Cambridge Univ. Press, 2013).
- Lorenz, R., Stalhandske, Z. & Fischer, E. M. Detection of a climate change signal in extreme heat, heat stress, and cold in Europe from observations. *Geophys. Res. Lett.* **46**, 8363–8374 (2019).
- Zwiers, F. W., Zhang, X. & Feng, Y. Anthropogenic influence on long returnperiod daily temperature extremes at regional scales. *J. Clim.* **24**, 881–892 (2011).
- Hoegh-Guldberg, O. et al. In *Global Warming of 1.5 °C. An IPCC Special Report on the Impacts of Global Warming of 1.5 °C Above Pre-industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change* (Intergovernmental Panel on Climate Change, 2018).
- Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Amer. Meteorol. Soc.* **93**, 485–498 (2012).
- Eyring, V. et al. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geoscientific Model Development (Online). 9. LLNL-JRNL-736881 (2016).
- Guo, Y., Gasparrini, A. & Armstrong, B. G. Coauthors. Heat wave and mortality: a multicountry, multicomunity study. *Environ. Health Perspect.* **125**, 087006 (2017).
- Vogel, E. et al. The effects of climate extremes on global agricultural yields. *Environ. Res. Lett.* **14**, 054010 (2019).
- Zuo, J. et al. Impacts of heat waves and corresponding measures: a review. *J. Clean. Prod.* **92**, 1–12 (2015).
- Vargas Zeppetello, L. R., Battisti, D. S. & Baker, M. B. The origin of soil moisture evaporation “regimes”. *J. Clim.* **32**, 6939–6960 (2019).
- Whan, K. et al. Impact of soil moisture on extreme maximum temperatures in Europe. *Weather Clim. Extremes* **9**, 57–67 (2015).
- Milly, P. C. D. & Dunne, K. A. Potential evapotranspiration and continental drying. *Nat. Clim. Change* **6**, 946–949 (2016).
- Schar, C. et al. The role of increasing temperature variability in European summer heatwaves. *Nature* **427**, 332–336 (2004).
- Stegehuis, A. I., Teuling, A. J., Ciais, P., Vautard, R. & Jung, M. Future European temperature change uncertainties reduced by using land heat flux observations. *Geophys. Res. Lett.* **40**, 2242–2245 (2013).
- Fischer, E. M., Rajczak, J. & Schär, C. Changes in European summer temperature variability revisited. *Geophys. Res. Lett.* **39**, L19702 (2012).
- Hersbach, H. et al. *Operational Global Reanalysis: Progress, Future Directions and Synergies with NWP* (European Centre for Medium Range Weather Forecasts, 2018).
- Hanlon, H., Hegerl, G. C., Tett, S. F. B. & Smith, D. Can a decadal forecasting system predict temperature extreme indices? *J. Clim.* **26**, 3728–3744 (2013).
- Seneviratne, S. I. et al. Investigating soil moisture–climate interactions in a changing climate: a review. *Earth Sci. Rev.* **99**, 125–161 (2010).
- Seneviratne, S. I. et al. Impact of soil moisture–climate feedbacks on CMIP5 projections: first results from the GLACE-CMIP5 experiment. *Geophys. Res. Lett.* **40**, 5212–5217 (2013).
- Miralles, D. G., Gentile, P., Seneviratne, S. I. & Teuling, A. J. Land–atmospheric feedbacks during droughts and heatwaves: state of the science and current challenges. *Ann. N. Y. Acad. Sci.* **1436**, 19 (2019).
- Vogel, Martha M., Zscheischler, Jakob & Seneviratne, Sonia I. Varying soil moisture–atmosphere feedbacks explain divergent temperature extremes and precipitation projections in central Europe. *Earth Syst. Dyn.* **9**, 1107–1125 (2018).
- Hall, A., Cox, P., Huntingford, C., & Klein, S. Progressing emergent constraints on future climate change. *Nat. Clim. Change* **9**, 269–278 (2019).
- Mitchell, Daniel et al. Half a degree additional warming, prognosis and projected impacts (HAPPI): background and experimental design. Geoscientific Model. Development. **10**, 571–583 (2017).
- Funk, C. et al. A high-resolution 1983–2016 Tmax climate data record based on infrared temperatures and stations by the climate hazard center. *J. Clim.* **32**, 5639–5658 (2019).
- Pal, J. S. & Eltahir, E. A. B. Future temperature in southwest Asia projected to exceed a threshold for human adaptability. *Nat. Clim. Change* **6**, 197–200, <https://doi.org/10.1038/NCLIMATE2833> (2017).
- Raymond, C., Matthews, T., & Horton, R. M. The emergence of heat and humidity too severe for human tolerance. *Sci. Adv.* **6**, <https://doi.org/10.1126/sciadv.aaw1838> (2020).
- van Vuuren, D. P. et al. The representative concentration pathways: an overview. *Clim. Change* **109**, 5–31 (2011).
- Gidden, M. et al. Global emissions pathways under different socioeconomic scenarios for use in CMIP6: a dataset of harmonized emissions trajectories through the end of the century. *Geosci. Model Dev. Discuss.* **12**, 1443–1475 (2019).
- King, A. D. et al. On the linearity of local and regional temperature changes from 1.5 °C to 2 °C of global warming. *J. Clim.* **31**, 7495–7514 (2018).
- Schurer, A. P. et al. Interpretations of the Paris climate target. *Nat. Geosci.* **11**, 220 (2018).
- Li, Z. & Yan, Z.-W. Homogenized daily mean/maximum/minimum temperature series for China from 1960–2008. *Atmos. Ocean. Sci. Lett.* **2**, 237–243 (2009).
- Cornes, R. C., van der Schrier, G., van den Besselaar, E. J. M. & Jones, P. D. An ensemble version of the E-OBS temperature and precipitation datasets. *J. Geophys. Res. Atmos.* **123**, 9391–9409 (2018).
- Weigel, A. P., Knutti, R., Liniger, M. A. & Appenzeller, C. Risks of model weighting in multimodel climate projections. *J. Clim.* **23**, 4175–4191 (2010).
- Collins, M. Still weighting to break the model democracy. *Geophys. Res. Lett.* **44**, 3328–3329 (2017).
- Lorenz, R. et al. Prospects and caveats of weighting climate models for summer maximum temperature projections over North America. *J. Geophys. Res. Atmos.* **123**, 4509–4526 (2018).
- Caldwell, P. M. et al. Statistical significance of climate sensitivity predictors obtained by data mining. *Geophys. Res. Lett.* **41**, 1803–1808 (2014).
- Berg, A. et al. Impact of soil moisture–atmosphere interactions on surface temperature distribution. *J. Clim.* **27**, 7976–7993 (2014).
- Brunner, L., Lorenz, R., Zumwald, M. & Knutti, R. Quantifying uncertainty in European climate projections using combined performance-independence weighting. *Environ. Res. Lett.* **14**, 124010 (2019).

## Acknowledgements

We acknowledge the E-OBS dataset from the EU-FP6 project UERRA (<http://www.uerra.eu>) and the Copernicus Climate Change Service, and the data providers in the ECA&D project (<https://www.ecad.eu>). This research used science gateway resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the US Department of Energy under Contract number DE-AC02-05CH11231. This research is funded by NERC grant award NE/S004661/1 EMERGENCE project.

## Author contributions

N.F. has performed all analyses related to this work and written the manuscript. G.H. has supervised the work and provided her expertise for output analysis. D.M. has contributed to the main text and results interpretation. M.C. has helped with the emergent constraint analysis and result discussions.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s43247-021-00094-x>.

**Correspondence** and requests for materials should be addressed to N.F.

**Peer review information** Primary handling editors: Heike Langenberg.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021